30/10/2026

EFAMRO and Esomar Consultation Response to the European Data Protection Board Guidelines on Web Scraping in the Context of Generative AI

EFAMRO the European Federation of Associations of Market Research Organisations. Founded in 1992, EFAMRO represents the interests of market, opinion and social research in Europe. Its members are national trade associations for research businesses.[1]

Esomar the worldwide, not-for-profit membership organisation committed to empowering and connecting the research, insights, and analytics community since 1947. Today, Esomar represents and provides connections among 50,000+ individuals and 750+ companies in 130+ countries.[2]

EFAMRO and Esomar represent the research and insights sector, accounting for a reported annual turnover of €34.87 billion in Europe.

1. About Market, Opinion and Social Research

1.1. Market research is comprised of all forms of market, opinion, and social research (herein referred to as "Market Research").

1.2. Market Research is the systematic gathering, analysis and interpretation of information about individuals or organisations. It uses the statistical and/or analytical methods and techniques of the applied social, behavioural, data and other sciences to generate insight and support decision making by providers of goods and services, governments, non-profit organisations and the general public.

1.3. Market Research stands at the heart of well-informed commercial, social and political decisions. Its purpose is to deliver information and insights about people's behaviour, needs and attitudes to inform decision making by providers of goods and services, governments, individuals, and society at large. Insight into what makes the success of a product, business initiative, consumer behaviour or government policy strategy is often the hidden – yet defining – factor that determines success and failure. It is our sector that provides the deeper intelligence needed for our world today by representing citizens' authentic voices, from all levels of society, in an unbiased and representative way by applying the tenets of statistical and scientific methods.

1.4. Market Research associations apply a comprehensive framework for self-regulation to ensure that researchers meet their ethical, professional, and legal responsibilities to the individuals whose data they use in research and to the clients and organisations which commission research. Members subscribe to self-regulation schemes that protect respondents' and participants' rights.

---

[1] For more about EFAMRO see: https://efamro.eu/

[2] For more about Esomar see: https://esomar.org

2. Purpose of our Response

2.1. EFAMRO and Esomar welcome the European Data Protection Board's work to clarify the application of the General Data Protection Regulation (GDPR) to the collection and use of publicly accessible online data in the development and use of generative artificial intelligence (AI). Clear, practical guidance is important both to protect individuals and to support responsible data use across different contexts.

2.2. The Guidelines appropriately emphasise the GDPR principles of transparency, accountability, data minimisation and the protection of individuals' rights and freedoms. Our associations particularly welcome the recognition that web scraping activities vary significantly in purpose, scope, risk and governance models and that assessments must remain context specific.

2.3. However, we are concerned that the draft Guidelines focus primarily on indiscriminate scraping for generative AI training without sufficiently distinguishing these activities from other lawful data-driven practices, such as those typical of the market, social and opinion research. Research organisations analyse publicly available online content to understand behaviour, attitudes and societal trends, using defined methods to answer specific research questions. Collection is generally limited to relevant information, and outputs are typically statistical or aggregated rather than directed at individuals. These differences may have a different effect on the assessment of necessity, proportionality, reasonable expectations, transparency obligations and legitimate interests.

2.4. We therefore recommend a number of clarifications to ensure that the final Guidelines remain proportionate, risk-based and applicable across legitimate research activities undertaken in Europe. Our comments below focus on clarifications that would help protect data subjects while ensuring that legitimate market, social and opinion research can be conducted responsibly.

3. Recognition of the distinction between research and generative AI training

3.1. Our associations welcome the distinction drawn by the EDPB between targeted and untargeted web scraping. However, the current drafting fails to explain more fully how this distinction affects the assessment of necessity, reasonable expectations, risk and legitimate interest.

3.2. A defined research project will ordinarily have a specific research question, a defined population or topic, documented inclusion and exclusion criteria, a limited collection period, quality controls and a planned analytical output. These features can materially limit the nature and consequences of the processing compared with unrestricted crawling designed to build an AI training dataset.

3.3. Therefore, we recommend that the final Guidelines expressly identify these project characteristics as relevant factors in the assessment. The analysis should consider the purpose and social value of the project, the source context, the data collected, the duration and scope of collection, the safeguards applied, and whether the outputs are aggregate or are used to make decisions about identifiable individuals.

4. Apply a contextual and proportionate legitimate-interests balancing test

4.1. We welcome the EDPB's confirmation that legitimate interest may, depending on the circumstances, provide an appropriate legal basis for web scraping activities where GDPR requirements are met. Our associations agree that controllers should be required to demonstrate necessity and conduct a careful balancing test that takes account of the data protection and privacy impacts on data subjects.

4.2. At the same time, we recommend that the Guidelines avoid treating the scale of data collection in itself as a determinative for legitimate interests. While scale is definitely relevant, it should be considered

alongside other contextual factors, including the purpose of processing, the source context, whether data were deliberately made public by the individual, the sensitivity of data, the duration of collection, the likelihood of identification, access controls, retention, and whether the activity produces only statistical or anonymous results. Research conducted to produce aggregate insights presents a significantly lower risk to individuals than processing designed to infer, predict or influence individual behaviour.

4.3. We therefore recommend that the final Guidelines emphasise that scale alone should not determine the outcome of a legitimate interest assessment and that the overall purpose, safeguards and likely impact on data subjects remain central factors.

4.4. Moreover, the Guidelines correctly identify reasonable expectations as an important component of the balancing test. However, reasonable expectations should be assessed with sufficient granularity. An individual may not reasonably expect an online post to be incorporated into an AI model capable of broad and unforeseen reuse. At the same time, the same individual may expect public comments, reviews or political discussion to be analysed in aggregate in order to understand attitudes, improve services monitor trends or inform policymaking.

4.5. The final Guidelines should therefore recognise that reasonable expectations depend not only on access to public information but also on the nature of the processing, the purpose pursued, whether outputs relate to individuals or groups and the likely consequences for the data subject.

5. Clarify controller, processor and downstream responsibilities at each stage

5.1. The Guidelines' discussion of the controller, processor and joint-controller roles is useful, in particular that the qualification of the organisations involved in each processing should be analysed on a case-by-case basis. However, we note that the legal and technical consequences of role attribution differ across collection, cleaning, structuring, model pre-training, fine-tuning, evaluation, retrieval-augmented generation, ordinary deployment and use of outputs.

5.2. The final Guidelines should therefore assess each processing operation separately and clarify that downstream users do not automatically inherit responsibility for every upstream collection activity. Responsibilities should be allocated in light of each actor's actual role, purposes, knowledge, influence, access to personal data and realistic ability to implement safeguards.

6. Provide practical and proportionate transparency guidance

6.1. Our associations support the EDPB's acknowledgement that individual notification under Article 14 GDPR may, in some situations, prove impossible or involve disproportionate effort. This issue is particularly relevant in the context of research that involves the collection and analysis of large volumes of publicly available data. In such circumstances, requiring individual notifications may present substantial practical difficulties while delivering limited additional transparency benefits.

6.2. We welcome the Guidelines' recognition in Clause 27 that Article 14(5)(b) GDPR may be relevant in the context of scientific research and statistical purposes. However, the draft text would benefit from greater clarity regarding the extent to which market, opinion and social research activities fall within this framework. Although Recital 159 GDPR and previous EDPB guidance encourage a broad interpretation of scientific research, the Guidelines do not clearly address whether research undertaken by commercial organisations, such as research agencies or insights providers, may benefit from the same considerations where the activity is conducted in accordance with recognised methodological, ethical and professional standards.

6.3. Greater clarity on this issue would provide important legal certainty for organisations conducting legitimate research while remaining fully consistent with the GDPR's objective of ensuring transparency through appropriate and proportionate means.

6.4. In relation to transparency obligations, we further encourage the draft Guidelines' recognition that alternative transparency mechanisms may be appropriate, particularly where processing is conducted for research, statistical or analytical purposes. In particular, in the context of a research project, a layered approach may provide more effective transparency. We therefore recommend that the Guidelines expressly acknowledge the use of layered and sector-level transparency mechanisms, including centrally accessible public notices supplemented, where appropriate, by project-specific information regarding individual research initiatives.

6.5. Greater clarity would also be helpful regarding the level of detail that controllers are expected to provide when relying on alternative transparency measures under Article 14 GDPR. In practice, organisations would benefit from further guidance on the extent to which transparency information should describe source categories, methods of collection, collection periods, processing purposes, categories of recipients, the use of AI systems or models, retention periods, applicable safeguards and the rights available to data subjects.

6.6. Finally, we encourage the recognition that, in limited and carefully justified circumstances, the immediate provision of certain information to data subjects may compromise the methodology of a research project. For instance, this issue may arise in blinded studies or observational research projects, or other forms of research where advance disclosure of specific methodological details could influence participants' behaviour and materially affect the quality and reliability of the findings. In such situations, a narrowly tailored delay or limitation of particular information may be appropriate, provided that the processing remains subject to appropriate safeguards, that the restriction is strictly necessary and proportionate, and that relevant information is made available to data subjects at the earliest appropriate opportunity.

6.7. Such an approach would be consistent with EDPB's Guidelines on processing of personal data for scientific research purposes, which state that where "the provision of information to data subjects may introduce bias to the research or otherwise negatively affect the quality of the personal data that is collected […], it may be justified to carry out so-called covert research, where it is allowed to not inform data subjects individually".

7. Clarify the treatment of special-category data, including incidental collection

7.1. We welcome the EDPB's effort to address the incidental collection of special category data and to explain the relevance of the Court's reasoning in GC & Others. However, further clarification is needed regarding circumstances in which market, social and opinion research may qualify as scientific research or statistical processing under Article 89 GDPR and Article 9(2)(j) GDPR.

7.2. Publicly available online data may reveal or permit inferences about health, political opinions, religious beliefs, trade-union membership and other special categories of personal data. Such information may be collected incidentally, or it may be central to legitimate, carefully designed research into health, elections, discrimination or social policy.

7.3. We therefore recommend that the final Guidelines distinguish between intentional and incidental collection, between information made public by the individual and information posted by third parties, and between directly observed information and inferred characteristics.

7.4. We further ask the EDPB to clarify, under Clause 67, whether market, social and opinion research may qualify as scientific research for processing of special-category data. Where the conditions for Article

9(2)(j) GDPR and Article 89 GDPR are satisfied, controllers may intentionally process special category data for research purposes, subject to an appropriate Article 6 legal basis, applicable Union or Member State law, necessity and suitable safeguards. The Guidelines should explain this potential route alongside the discussion of incidental and residual special-category data, including the reasoning referred to in GC & Others, and consistently with the EDPB's separate guidance on scientific research. A risk-based assessment should not imply that the presence of special-category data necessarily makes all research processing impermissible.

8. Interpret technical restrictions according to their purpose and scope

8.1. Clauses 37 and 64 discuss robots.txt, ai.txt and CAPTCHA as indications that a website opposes scraping. These mechanisms may have different purposes and scopes: some target particular crawlers or AI training, while others restrict automated access more broadly. Treating these mechanisms as equivalent risks overstating the scope of a website operator's intentions and may unnecessarily restrict legitimate, targeted research activities. We therefore recommend that the final Guidelines distinguish more clearly between these signals and clarify they should be interpreted and respected according to their specific purpose and scope.

8.2. At the same time, compliance with a technical restriction should not be regarded as the sole determinant of the lawfulness of scraping. Contractual terms, intellectual-property rights, data-owner permissions, data-subject rights and other applicable legal obligations remain relevant.

9. Clarify the assessment of the principle of accuracy

9.1. Our associations welcome the EDPB's recommendations to use reliable sources, timestamp material and validate data to support compliance with the GDPR principle of accuracy.

9.2. At the same time, we encourage the EDPB to recognise that the concept of accuracy may require a more nuanced application in the context of opinion, social and behavioural research. In many research contexts the phenomenon being studied may include inaccurate beliefs, rumours, misinformation, hostile discourse or subjective experiences. In such cases, accuracy should not be understood solely as the factual correctness of the underlying content. Rather, accuracy may require that the information is faithfully recorded and analysed in a manner that preserves the context in which it was expressed. A dataset that accurately captures a widespread misunderstanding or inaccurate belief may nevertheless constitute an accurate record of the social phenomenon under investigation.

9.3. To support the principle of accuracy, the recommendations on source validation (clause 42) and data cleaning (clause 11) could be expanded to encourage proportionate provenance checks, identification and labelling of AI-generated content where reasonably feasible, and documentation of any uncertainty regarding source authenticity.

9.4. The key consideration should therefore not be whether content is reliable or unreliable, but whether the chosen approach is appropriate for the intended purpose and whether adequate measures have been implemented to address risks relating to data quality, representativeness, provenance and accuracy. This is particularly relevant from a data protection perspective where distorted or low-quality training data may contribute to inaccurate outputs concerning identifiable individuals.

10. Apply data minimisation in a way that preserves research validity

10.1. Our associations welcome how the Guidelines appropriately emphasise data minimisation. In the context of market, social and opinion research, data minimisation should be assessed in relation to a defined and legitimate research purpose, while taking account of validity, representativeness, reproducibility and the need to identify bias. Excluding all websites that oppose scraping, or entire

categories of websites, may create systematic biases, distort longitudinal evidence and prevent meaningful subgroup testing. We therefore encourage the EDPB to recognise that data minimisation assessments should consider not only privacy risks but also the potential implications for the validity and accuracy of research findings.

10.2. Similarly, synthetic data should not be presented as an inherently preferable alternative to observed data. While synthetic data may reduce some privacy risks, support testing or fill specific gaps, it is not inherently anonymous, unbiased, representative or current, and it may reproduce patterns from the personal data used to generate it. For research into actual attitudes and behaviour, synthetic data may not be a methodologically valid substitute for observed evidence. It should therefore be presented as one possible safeguard, subject to disclosure-risk and utility testing, rather than a default alternative whose non-use indicates non-compliance.

10.3. Synthetic data may reproduce individual variable distributions without adequately preserving relationships between variables, minority groups or unusual events. Adding synthetic records does not, by itself, add independent observations of the population. Uncertainty estimates should reflect the generation process, dependence on source data and model assumptions. Where the Guidelines suggest an alternative should be equally effective, synthetic data should be treated as such only where evidence supports its effectiveness for the specific purpose.

11. Recognise Sector-Specific Safeguards and Professional Governance

11.1. Our associations recommend that the final Guidelines give greater recognition to professional research codes of conduct, sector-specific ethical standards and self-regulatory frameworks as relevant components of an organisation's broader governance and safeguarding arrangements.

11.2. Market, opinion and social research operates within an established professional framework, including the ICC/Esomar International Code on Market, Opinion and Social Research and Data Analytics, alongside applicable national codes and professional standards.

11.3. These frameworks address confidentiality, participant protection, responsible data use and the ethical risks that may arise when research information is combined, interpreted or used in ways that could affect individuals.

11.4. This is particularly relevant to Clauses 65 and 66 regarding the introduction of mitigating measures aimed at limiting the impact of the processing on data subjects. The final Guidelines should recognise sector safeguards such as the ICC/Esomar International Code and applicable national codes, as relevant when assessing mitigating measures and can contribute to accountability, transparency and risk reduction beyond the legal minimum requirements.

12. Conclusion

12.1. EFAMRO and Esomar encourage the EDPB to refine the Guidelines so that they remain firmly grounded in the GDPR while reflecting the material differences between unrestricted web scraping for generative AI development and targeted, safeguarded research. A proportionate, context-sensitive approach can protect individuals' rights while preserving the ability of market, social and opinion research to generate reliable aggregate evidence for organisations, policymakers and society.

12.2. EFAMRO and Esomar welcome the opportunity to assist the European Data Protection Board in ensuring the Guidelines 03/2026 on web scraping in the context of generative AI are proportionate, fit-for-purpose, targeted and contextually appropriate.

To contact us for more information:

- Kaleke Kolawole, Head of Policy: kaleke.kolawole@efamro.eu
- Federica Pizzuti, Public and Government Affairs Specialist, federica.pizzuti@esomar.org